White Manipulation in Judgment Aggregation

Davide Grossi [a] Gabriella Pigozzi [b] Marija Slavkovik [b]

[a] *Universiteit van Amsterdam, P.O. Box 94242, 1090 GE Amsterdam.*
E-mail: d.grossi@uva.nl
[b] *University of Luxembourg, 6 Rue Richard Coudenhove–Kalergi, L-1359 Luxembourg.*
E-mail: {gabriella.pigozzi,marija.slavkokvik}@uni.lu

Abstract

Distributive systems consisting of autonomous and intelligent components need to be able to reason and make decisions based on the information these components share. Judgment aggregation investigates how individual judgments on logically connected propositions can be aggregated into a collective judgment on the same propositions. It is the case that seemingly reasonable aggregation procedures may force the group to hold an inconsistent judgment set. What happens when the agents realize that the group outcome will be inconsistent? We claim that, in order to avoid an untenable collective outcome, individuals may prefer to declare a non-truthful, less preferred judgment set. Thus, the prospect of an individual trying to manipulate the social outcome by submitting an insincere judgment set is turned from being an undesirable to a "virtuous" (or white) manipulation. We define white manipulation and present the initial study of it as a coordinated action of the whole group.

1 Introduction

The increase of distribution in intelligent systems calls for increase in intelligence and autonomy on the part of the distributed elements. Furthermore, following the trend of assets reuse in computer science, it has to be taken into account that the constituting elements will be part of more than one distributed system at a time. Consequently, the distributed elements can no longer be satisfactorily modeled as inert, but need to be considered and reasoned about as artificial autonomous agents. To maintain the autonomy of the distributed system as a whole, the constituting agents will inevitably face problems of collective reasoning and collective decision-making. Such problems are native to social choice theory. Thus, the requirement emerges to study social theoretic problems from the computer science perspective, as well as to modify the known solutions to solutions that solve artificial agents problems.

In our work we focus on the sub-discipline of social choice known as judgment aggregation. Judgment aggregation [5, 6] studies how groups of agents aggregate individual judgments on logically connected propositions into a collective judgment on the same propositions. Consider as an example a group of three database administrators (DBAs) which need to decide collectively whether to make an existing agent X into a fourth DBA. All DBAs need to be synchronized in their activities in order to maintain the integrity of the database they manage. Consequently, the addition of a new DBA, apart from reducing the workload of the existing DBAs, increases the communication and interaction costs. Since the administrator privileges include the ability to make security sensitive changes, any new DBA needs to be trustworthy. The DBAs necessarily follow the decision rule: a new DBA is created if and only if the existing DBAs judge that they can not handle the work load, and the candidate user is judged trustworthy.

Assume that each DBA expresses yes/no opinions (judgments) on the propositions (the workload too much for the existing DBAs, X is trustworthy), and the corresponding conclusion (make the candidate X a fourth DBA) according to Table 1.

The group may have to endorse an inconsistent position if the groups opinion is taken to be the majority view on each of the issues separately. This is an instance of the so-called *discursive paradox* [8].[1]

[1] It is important to mention that the problem of aggregating individual judgments is not restricted to majority voting, but it applies to all aggregation procedures satisfying some seemingly desirable conditions.

	DBA1	DBA2	DBA3	Majority
workload too much for current DBAs	*yes*	*yes*	*no*	*yes*
X is trustworthy	*yes*	*no*	*yes*	*yes*
make(X,DBA,4)	*yes*	*no*	*no*	*no*

Table 1: Creating a new DBA example. The candidate X is assigned a DBA status if and only if the current DBAs can not handle the workload and X is trustworthy.

The discursive paradox shows that the three DBAs are in an impasse. As a group, they cannot claim to make X into a DBA despite that there is a majority voting yes for both premises. The three DBAs need to find an agreement on whether to create the new DBA. Additionally, the DBAs need to provide consistent reasons that support their decision. The reasons are necessary, since the DBAs have to account for their final decision not only to the database users but also to the database owner. It is also important to consider that the deliberation can not be prolonged indefinitely since the database may need to be locked during the deliberation period.

Let us assume the judgments revealed in Table 1 are sincere and represent the definite answer on what each DBA holds to be true. It would still be reasonable for the DBAs to discuss the individual opinions, try to find a collective agreement and to agree on a collective consistent outcome. One possibility to achieve this is through presenting insincere judgments. Generally the case when an agent lies in declaring information is known as manipulation.

In this paper we are interested at investigating a "positive" notion of manipulation, which arises when agents are available to adjust their honest judgments in order to avoid that the group falls into an impasse. The manipulability of social choice decision rules is usually considered an undesirable property. If a decision rule is manipulable an agent may, upon learning the preferences of the other agents, misrepresent his input to ensure a social choice decision in his favor. However, an agent's incentive may not be to have his own input selected as the collective choice. Instead, an agent may have an "unselfish" incentive to improve the possibility of reaching a swift social choice decision. Like the DBAs in the example, agents may have a justified incentive to avoid the impasse as the social choice, and manipulate the social choice inputs to this end. The agents are willing to deliberately fallback into declaring a less preferred input to ensure that a decision can be made. Hence, we analyze how the agents can achieve two goals simultaneously: to ensure a rational group decision and, at the same time, to do so by diverging the slightest possible from their own sincere judgments. In colloquial contexts, "white lie" is the term used to indicate a well-intended lie told with the purpose of being diplomatic and/or avoiding to hurt others. Here we will use *white manipulation* to term the scenarios in which benevolent agents manipulate group decisions in order to avoid an impasse. The aim of the work here presented is to provide a preliminary investigation in white manipulation.

As agreement reaching procedure we consider *fallback bargaining* [1]. In fallback bargaining the individuals indicate a preference ordering over the alternatives. If an agreement is not found at the level of the most preferred alternative, the individuals fall back to less and less preferred alternatives until an agreement is reached. Group members use fallback bargaining to negotiate their manipulations and to escape an inconsistent group outcome.

The paper will proceed as follows. First, in Section 2, we introduce the judgment aggregation framework. In Section 3 we extend such framework by introducing agents' preferences. This is the first essential step in order to enable considerations over white manipulability within the aggregation problem. In Section 4 we present fallback bargaining and we show how fallback bargaining can be employed on judgment set profiles. Due to space limitations, the proofs of the theorems are placed in the appendix.

2 Judgment Aggregation

The present section introduces the framework of judgment aggregation along with the notation used in this paper. For a general overview of the field the reader is referred to [7].

In judgment aggregation, the issues upon which the agents are called to express acceptance or rejection consist of logical formulae. Such a set of issues is called an *agenda*. In this paper we work with the standard setting of judgment aggregation [6] based on propositional logic. Let $\mathcal{L}$ be a propositional language, and let $l(\mathcal{L})$ be its signature, i.e., the set of atoms upon which $\mathcal{L}$ is built. We can now define the notion of agenda.

Definition 2.1 (Agenda $\mathcal{A}$)**.** The agenda $\mathcal{A} \subseteq \mathcal{L}$ is a finite set of formulas of $\mathcal{L}$ such that $l(\mathcal{A}) \subseteq \mathcal{A}$.

Sometimes it will be convenient for us to think of agendas as m-tuples of issues, where $m = |\mathcal{A}|$, that is, to assign a number j such that $1 \leq j \leq |\mathcal{A}|$ refers to j-th element in $\mathcal{A}$.

Example 2.1. For the DBA example from Table 1, the corresponding agenda for the group is $\mathcal{A} = \{p_1, p_2, p_3\}$, with $l(\mathcal{A}) = \{p_1, p_2, p_3\}$.

Each agenda comes with a set of logical constraints, also called rules, which make explicit what logical relations occur between the issues in the agenda.

Definition 2.2 (Agenda constraints $\mathcal{R}^{\mathcal{A}}$)**.** The set $\mathcal{R}^{\mathcal{A}} \subseteq \mathcal{L}$ is a finite, logically consistent, set of formulas such that $l(\mathcal{R}^{\mathcal{A}}) \subseteq l(\mathcal{A})$.

Given an agenda $\mathcal{A}$ and a set of rules $\mathcal{R}^{\mathcal{A}}$ for $\mathcal{A}$, each agent casts an acceptance or rejection vote on each issue in $\mathcal{A}$ by keeping consistency with respect to $\mathcal{R}^{\mathcal{A}}$. The notion of *judgment set* can now be conveniently defined as follows.

Definition 2.3 (Judgment set φ)**.** A judgment set φ is a map $\varphi : \mathcal{A} \longrightarrow \{0, 1\}$ which satisfies the formulas in $\mathcal{R}^{\mathcal{A}}$ according to the semantics of propositional logic. The set of all judgment sets for $\mathcal{A}$ which satisfy the formulas in $\mathcal{R}^{\mathcal{A}}$ is called *domain* and is denoted $\Phi(\mathcal{A}, \mathcal{R}^{\mathcal{A}})$.[2]

Hence, a judgment set is nothing but a $\mathcal{R}^{\mathcal{A}}$-consistent assignment of 1 (acceptance) or 0 (rejection) to the elements of $\mathcal{A}$. Finally, we can define when two judgment sets agree on an agenda item [3]. Just like we can think of agendas as n-tuples of issues, we can conceive judgment sets as n-tuples of values in $\{0, 1\}$, where $n = |\mathcal{A}|$. In this case, the value that φ assigns to the j^{th} issue in $\mathcal{A}$ is termed *j-projection* of φ and it is denoted $\pi_j(\varphi)$.

Definition 2.4. Let φ and φ' be judgment sets for an agenda $\mathcal{A}$ which satisfy the formulas in $\mathcal{R}^{\mathcal{A}}$. We say that judgment set φ agrees with judgment set φ' on the j^{th} agenda item if $\pi_j(\varphi) = \pi_j(\varphi')$. The judgment sets φ and φ' are otherwise said to disagree on the j^{th} agenda item.

Judgment aggregation investigates how to assign a collective judgment set to a profile of individual judgment sets, representing the view that each agent holds on the issues in the agenda.

Definition 2.5 (Judgment profile ω)**.** Let $N = [1, n]$ be a finite set of n agents, $\mathcal{A}$ an agenda, and $\mathcal{R}^{\mathcal{A}}$ a set of rules for $\mathcal{A}$. A judgment profile ω for the group N is a n-tuple $\omega = (\varphi_i)_{i \in N}$ such that φ_i satisfies $\mathcal{R}^{\mathcal{A}}$. The set of all possible judgment profiles ω for a given group N, agenda $\mathcal{A}$ and set of rules $\mathcal{R}^{\mathcal{A}}$ is called *universe* and is denoted by $\Omega(N, \mathcal{A}, \mathcal{R}^{\mathcal{A}})$.[3]

For later use, we also define what it means for one profile to be an i-variant of another.

Definition 2.6 (i-variant)**.** Consider $\omega', \omega'' \in \Omega$. We say that ω' is an i-variant of ω'' if and only if ω' coincides with ω'' on all judgment sets except the i^{th} one.

The final ingredient of a judgment aggregation framework is the *aggregation function*, that assigns collective judgment sets to judgment profiles.

Definition 2.7 (Aggregation function f)**.** Let N be a set of n agents, $\mathcal{A}$ an agenda, and $\mathcal{R}^{\mathcal{A}}$ its set of rules. An aggregation function is a function $f : \Omega(N, \mathcal{A}, \mathcal{R}^{\mathcal{A}}) \longrightarrow \Phi(\mathcal{A}, \mathcal{R}^{\mathcal{A}})^{\lightning}$ where $\Phi(\mathcal{A}, \mathcal{R}^{\mathcal{A}})^{\lightning} = \Phi(\mathcal{A}, \mathcal{R}^{\mathcal{A}}) \cup \{\lightning\}$.

In other words, an aggregation function is a function from the universe Ω to $\Phi^{\lightning}$ that is, the domain Φ plus the element denoted $\lightning$. Intuitively, $\lightning$ denotes all those outcomes that are assignments not satisfying the constraints holding on the agenda, i.e. a *collectively irrational* outcome. We take that for any j, $\mathcal{A}$ and $\mathcal{R}^{\mathcal{A}}$ it holds $\pi_j(\lightning) = \lightning$.

We now define a family of aggregation functions – the *q-rule*. The intuition behind such functions is simple, the inspiration pulling from the q-rule in voting. A threshold q is fixed, for all the proposition, such that $\lfloor \frac{n}{2} \rfloor + 1 \leq q \leq n$, where n is the number of agents. A proposition is collectively accepted if and only if the number of group members accepting it is at least equal to q.

[2] In what follows we often drop the argument $(\mathcal{A}, \mathcal{R}^{\mathcal{A}})$ for the ease of notation.

[3] In the following, we often omit the argument $(N, \mathcal{A}, \mathcal{R}^{\mathcal{A}})$.

Definition 2.8 (q-rule)**.** Let $\omega = (\varphi_1, \ldots, \varphi_n)$ be any judgment profile for a set of n agents N. We denote with $\omega^{j,0}$ and $\omega^{j,1}$ the subsets of ω that agree on the j^{th} agenda item and:
$\omega^{j,0}$ contains all the judgment sets for which that agenda item is $p_j(\varphi_i) = 0$.
$\omega^{j,1}$ contains all the judgment sets for which that agenda item is $p_j(\varphi_i) = 1$.

The q-rule aggregation functions f^q are defined as follows:

$$f^q(\omega) = \begin{cases} \varphi & \text{if } \varphi \text{ is a judgment set and for each } \pi_j(\varphi) \text{ it holds that either} \\ & \pi_j(\varphi) = 1, \mid \omega^{j,1} \mid > \mid \omega^{j,0} \mid \text{ and } \mid \omega^{j,1} \mid \geq q \text{ or} \\ & \pi_j(\varphi) = 0, \mid \omega^{j,0} \mid > \mid \omega^{j,1} \mid \text{ and } \mid \omega^{j,0} \mid \geq q \\ \lightning & \text{otherwise} \end{cases} \tag{1}$$

When $q = n$ we speak of *unanimous rule* and we denote f^{un}. When $q = \lfloor \frac{n}{2} \rfloor + 1$ we speak of *propositionwise majority rule*, which we denote with f^{maj}.

Example 2.2. Consider the profile in Table 1 and call it ω. We have that: $f^{maj}(\omega) = \lightning$ and $f^{un}(\omega) = \lightning$. Notice that, the fact that $f^{maj}(\omega) = \lightning$ while the set of judgments constructed obtained by propositionwise majority voting is defined, i.e. $(1, 1, 0)$, captures the discursive dilemma, that is the existence of an 'irrational' majority.

3 Preferences over Aggregation Outcomes

Agents faced with a group decision usually hold preferences over the outcomes of the decision process. In our setting, this is to say that individuals have preferences over the set of possible outcomes $\Phi^{\lightning} = \Phi \cup \{\lightning\}$. We can define the notions of preference and preference profile in the usual way.

Definition 3.1 (Preferences and preference profiles)**.** A preference over $\Phi^{\lightning}$ is a total pre-order on $\Phi^{\lightning}$. Given a set N of n agents, a preference profile for N over $\Phi^{\lightning}$ is a n-tuple of total pre-orders on $\Phi^{\lightning}$.

The present section shows how agents preferences can be read off a given judgment profile (and a judgment aggregation function) by assigning to each agent a preference ordering over $\Phi^{\lightning}$ for every judgment profile ω in Ω. The intuition behind this extension of the judgment aggregation framework is that a profile ω can naturally be interpreted as a tuple which represents, for each agent i, the judgment set that i most prefers. To obtain the full preference ordering of each agent, we then apply a specific notion of *distance* between judgment sets. Our is a development of the *top-respecting preferences* as in [3] using scoring functions.

Intuitively, if φ is the judgment set that is most preferred by i, then φ' will be strictly preferred to φ'' if φ' differs from φ strictly less than φ'' does. In other words, given a judgment profile ω, we are interested in methods to extract total pre-orders over the set of all the judgment sets plus $\lightning$ (i.e., all the possible outcomes of an aggregation). Such a method is therefore a function $s : \Phi \longrightarrow (\Phi^{\lightning} \longrightarrow \mathbb{R})$ and is called a *scoring function*.We present one way of extracting the ordinal information we are interested in from a given judgment profile.

A Hamming distance measures, given two tuples (or strings) of values, the number of entries for which the values in the two strings differ. As shown in [9], this idea can be straightforwardly applied to judgment sets as they have been defined in Definition 2.3.

Definition 3.2 (Hamming distance $H(\varphi_1, \varphi_2)$)**.** Let $\mathcal{A}$ be an agenda, $\mathcal{R}^{\mathcal{A}}$ its set of constraints, and Φ the set of judgment sets. The Hamming distance $H : \Phi^2 \longrightarrow \mathbb{R}$ between two judgment sets $\varphi_1, \varphi_2 \in \Phi$ is defined as in Equation 2 with functions $d_i : \Phi^2 \longrightarrow \{0, 1\}$ defined in Equation 3.

$$H(\varphi_1, \varphi_2) = \sum_{i=1}^{|\mathcal{A}|} d_i(\varphi_1, \varphi_2) \tag{2}$$

$$d_i(\varphi_1, \varphi_2) = \begin{cases} 0 & \text{if } \pi_i(\varphi_1) = \pi_i(\varphi_2) \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

The Hamming distance H calculates how "far" φ_2 is from φ_1 by counting the numbers of formulas in $\mathcal{A}$ to which φ_2 assigns a different value than the one assigned by φ_1. In other words, judgment sets are just considered to be $|\mathcal{A}|$-tuples of values in $\{0, 1\}$ and H counts the numbers of different entries in two given tuples. It is easy to see that $H(\varphi, \varphi) = 0$, that $H(\varphi_1, \varphi_2) = H(\varphi_2, \varphi_1)$, and that the maximum possible pseudo-distance coincides with the cardinality $|\mathcal{A}|$ of the agenda.

However, the Hamming distance as defined in Definition 3.2 does not provide a way of measuring the distance of $\lightning$ from a given judgment set. This is because $\lightning$ is not a judgment set (Definition 2.3), and $H(\varphi_1, \varphi_2)$ is defined only over judgment sets. The way agents rank $\lightning$ will be kept variable. The next definition introduces a score based on H for any element in $\Phi^{\lightning}$.

Definition 3.3 (H-score)**.** Let $\mathcal{A}$ be an agenda, $\mathcal{R}^{\mathcal{A}}$ its set of constraints, and Φ the set of judgment sets. Take $\varphi \in \Phi$. The H-score function $HS_e : \Phi \longrightarrow (\Phi^{\lightning} \longrightarrow \mathbb{R})$ is defined as in Equation 4, where $\varphi \in \Phi$ and $e : \Phi \longrightarrow [1, |\mathcal{A}| + 1]$.

$$HS_e(\varphi)(\varphi') = \begin{cases} H(\varphi, \varphi') & \text{if } \varphi' \in \Phi \\ e(\varphi) & \text{otherwise} \end{cases} \tag{4}$$

The H-score is a function that, for each judgment set $\varphi \in \Phi$ and outcome in $\Phi^{\lightning}$, yields a score in $\mathbb{R}$. So, Definition 3.3 introduces a family of functions which all use the Hamming distance H to rank outcomes in $\Phi^{\lightning}$ given a judgment set φ, which is taken to be the most preferred one. The H-score makes use of an independent function e to provide the score of $\lightning$ as a function of φ. Intuitively, a function e tells us how $\lightning$ is ranked within the preferences of the agents. Depending on the chosen e, the inconsistent outcome $\lightning$ can for example obtain values between 0 and 1, thus becoming the second most preferred outcome, and between 1 and $|\mathcal{A}| + 1$, becoming instead the least preferred one.

These observations are made concrete in the formula below. Given a judgment set φ taken as the most preferred by agent i, it is then a matter of applying Definition 3.3 to obtain the preference ordering of i over the set $\Phi^{\lightning}$ of all judgment sets yielded by $HS_e(\varphi)$:

$$\varphi' \preceq^{HS_e(\varphi)} \varphi'' \quad \text{iff} \quad HS_e(\varphi)(\varphi'') \leq HS_e(\varphi)(\varphi') \tag{5}$$

where $\varphi', \varphi'' \in \Phi^{\lightning}$ and $\varphi \in \Phi$. That is to say, φ' is at most as preferred as φ'' if and only if φ' has extended Hamming distance (HS_e) from φ at least equal to the extended Hamming distance of φ'' from φ. It follows that $\preceq^{HS_e(\varphi)}$ is a total pre-order over $\Phi^{\lightning}$. When it is clear from which judgment set φ the preference is derived, we will use the lighter notation $\preceq^{HS_e}$.

It is now easy to see that, on the ground of Definition 3.1, each judgment profile ω univocally determines a profile of $\preceq^{HS_e}$-preferences. This is done in a natural way by applying function HS_e to each judgment set φ_i in $\omega = (\varphi_i)_{i \in N}$, thus obtaining a preference profile $(\preceq_i^{HS_e})_{i \in N}$, which we denote $HS_e(\omega)$. We show how this can be done in our running example.

Example 3.1. DBA example continued ($\preceq^{HS_e}$-profiles). Consider the profile $\omega = (\varphi_1, \varphi_2, \varphi_3)$ corresponding to Table 1. Take e to be such that $e(\varphi_1) = e(\varphi_2) = e(\varphi_3) = 4$, that is, all agents give the highest possible H-score to $\lightning$. We have the following preference profile $HS_e(\omega)$ where each preference is obtained by applying HS_e to the corresponding judgment set in ω:

$$\begin{aligned}
&(1,1,1) \succ_1^{HS_e} (1,0,0) \sim_1^{HS_e} (0,1,0) \succ_1^{HS_e} (0,0,0) \succ_1^{HS_e} \lightning \\
&(1,0,0) \succ_2^{HS_e} (0,0,0) \succ_2^{HS_e} (0,1,0) \sim_2^{HS_e} (1,1,1) \succ_2^{HS_e} \lightning \\
&(0,1,0) \succ_3^{HS_e} (0,0,0) \succ_3^{HS_e} (1,0,0) \sim_3^{HS_e} (1,1,1) \succ_3^{HS_e} \lightning
\end{aligned}$$

This profile depicts one of the possible decision-theoretic situations underlying the discursive dilemma. In fact, function e identifies precisely the variable factor in agents' preferences which can potentially lead an aggregation process to stall depending on how highly $\lightning$ is ranked within the agents' preferences.

Given a scoring function and a judgment aggregation function f, we can define a pre-order over profiles of judgment sets.

Definition 3.4. Let Ω be a set of all possible profiles ω of judgment sets over some agenda $\mathcal{A}$ and a set of rules $\mathcal{R}^{\mathcal{A}}$. Let f be a judgment aggregation function. Given the scoring function $HS_e(\varphi)$, we define a total pre-order over each $\omega \in \Omega$ as $\omega' \preceq^{HS_e(\varphi)} \omega''$ iff $HS_e(f(\omega))(f(\omega'')) \leq HS_e(f(\omega))(f(\omega'))$.

The H-score is one way to provide a distance-based construction of agents' preferences out of the most preferred judgment set. Other distance based scoring functions can be defined in a similar way. In fact the H-score is guaranteed to produce indifference among some judgment sets, which is not always desirable. At present, in the remaining of the paper we rely on the H-score to illustrate our work.

3.1 Preference aggregation in judgment aggregation

We have given examples of distance-based constructions of preference profiles from judgment profiles. Independently of the specific scoring function $s : \Phi \longrightarrow (\Phi^{\lightning} \longrightarrow \mathbb{R})$ at hand, it is now possible to include social choice theoretic concepts in the judgment aggregation framework. Given the universe Ω of a framework with N agents, function s yields the set of all judgment aggregation preference profiles over $\Phi^{\lightning}$, which

we denote $s(\Omega)$. Now the tuple $(N, \Phi^{\not\iota}, s(\Omega))$ fully specifies a preference aggregation problem [4] where N is the set of agents, $\Phi^{\not\iota}$ is the set of outcomes, and $s(\Omega)$ is the set of all judgment aggregation preference profiles.[4]

With agents' preferences obtained by means of a numerical scoring function s, the notions of utilitarian and egalitarian social welfare are straightforwardly applicable. In our case, for a given judgment profile $\omega = (\varphi_i)_{1 \leq i \leq n}$ and scoring function s, the utilitarian social welfare of an outcome $\varphi \in \Phi^{\not\iota}$ has to be understood as the sum of all the distances of φ from agents' judgments in ω, that is as in Equation 6.

The egalitarian social welfare of an outcome $\varphi \in \Phi^{\not\iota}$ is instead the welfare of the worst off agent, which in our case is the agent whose most preferred judgment set is the furthest from the one selected in the outcome. That is as in Equation 7.

$$USW_{s(\omega)}(\varphi) = \sum_{i=1}^{n} s(\varphi)(\varphi_i) \quad (6) \qquad ESW_{s(\omega)}(\varphi) = \texttt{max}\{s(\varphi)(\varphi_i) \mid \varphi_i \in \Phi\} \quad (7)$$

Formulas 6 and 7 above define social choice functions mapping the set of all preference profiles that are built via s and a judgment profile, to the set of outcomes $\Phi^{\not\iota}$. In other words, scoring functions s allow us to define voting rules for judgment aggregation, which are based on the consideration of agents' preferences. In social choice theory, a (non-resolute) voting rule picks, for any preference profile, a non-empty subset of the set of outcomes [10]. The definition of two examples of such functions follows.

Definition 3.5 (Utilitarian and egalitarian social choice in judgment aggregation)**.** Let N be a set of agents, Φ the set of possible judgment sets given $\mathcal{A}$ and $\mathcal{R}^{\mathcal{A}}$, and s a scoring function.

- The utilitarian social choice function $usc : s(\Omega) \longrightarrow \Phi^{\not\iota}$ is defined as follows:
  $usc(s(\omega)) = \{\varphi \in \Phi^{\not\iota} \mid USW_{s(\omega)}(\varphi) = \texttt{min}(\{USW_{s(\omega)}(\varphi') \mid \varphi' \in \Phi^{\not\iota}\})\}$
- The egalitarian social choice function $esc : s(\Omega) \longrightarrow \Phi^{\not\iota}$ is defined as follows:
  $esc(s(\omega)) = \{\varphi \in \Phi^{\not\iota} \mid ESW_{s(\omega)}(\varphi) = \texttt{min}(\{USW_{s(\omega)}(\varphi') \mid \varphi' \in \Phi^{\not\iota}\})\}$

Intuitively, the utilitarian function maximizes the sum of the individual welfare of the whole group (by minimizing the sum of the distance of the outcome from the top choices of each agent), and the egalitarian one maximizes, instead, the welfare of the agent which is worst off.

In this section we have seen how notions from social welfare theory can be applied within a judgment aggregation setting. This will allow us to define the phenomenon of "white manipulation" we sketched in the introduction.

3.2 White manipulation

An agent manipulates the judgment aggregation if he submits a judgment set which is not his true choice of valuation for the elements of the agenda (truthful judgment set). The Definition 3.6 we present corresponds to the definition of non-manipulability presented in [3].

Definition 3.6 (Manipulability)**.** Let s be a scoring function. An aggregation function f is manipulable if and only if there exists a judgment profile $\omega \in \Omega$ and an agent i such that $f(\omega) \prec_i^s f(\omega')$, where $\omega' \in \Omega$ is some i-variant of ω.

Definition 3.7 (White manipulability)**.** Let $\mathcal{SW}$ be a social welfare function and s a scoring function. An aggregation function f is white manipulable if and only if there exists an agent i and a judgment profile $\omega \in \Omega$ such that $f(\omega) \prec_i^s f(\omega')$ and $\mathcal{SW}(f(\omega)) < \mathcal{SW}(f(\omega'))$, where $\omega' \in \Omega$ is some i-variant of ω.

For example, the social welfare function $\mathcal{SW}$ can be one of the utilitarian or egalitarian functions we defined in Equation 6 or Equation 7 [5].

A full analysis of the characteristics which make a judgment aggregation function white manipulable is left for future work. Here we propose one method, fallback bargaining, in which white manipulation can be conducted on a group level.

[4] It is worth noticing that, depending on s, $s(\Omega)$ might be a strict subset of the set of all total pre-orders that can be built on $\Phi^{\not\iota}$. In such cases, the aggregation does not satisfy the so-called *universal domain* [7] condition. This is the case, for instance, of the H-score function (Definition 3.3).

[5] Note that in this case the definition should read $USW(f(\omega')) < USW(f(\omega))$ since the higher the score, the less preferred an element from $\Phi^{\not\iota}$ is.

4 Fallback Bargaining

White manipulation can be conducted by one benevolent agent under the assumption that the rest of the agents submit their honest judgment sets. Since the goal of the white manipulation is "for the benefit of all", it is of interest to view the white manipulation as a group coordinated activity. Additionally, if two agents uncoordinatedly manipulate the judgment aggregation, the outcome may still not be the desired one. We now show how, by using fallback bargaining, can the group negotiate their "white lies" with the goal of selecting a profile that, under a known judgment aggregation function, yields an outcome with a lower USW than the truthful profile. In particular we are interested in white manipulation when the outcome of the judgment aggregation function is $\lightning$.

Fallback bargaining is an agreement reaching procedure presented in [1, 2]. The procedure accepts as input a set of k possible alternatives K and, for each agent in the set of n agents N, a total preference ordering over the elements of K. Using the preferences of the agents, a matrix $M(m_{ij})$ is created where each row represents an agent and each column the preference orderings of the agents. Thus, for example, element m_{32} holds the second most preferred alternative of the second agent. The agent i's ranking for alternative x is denoted with $j_i^M(x)$. The enumeration of the columns is referred to as d standing for *depth of preference* . The r-approval bargaining, fully described in [1], is in fact the process of creating sets $CS_d^r(M) = \{x \in K : |\{i : j_i^M(x) \leq d\}| \geq r\}$ starting from $d = 1$ until $CS_d^r(M) \neq \varnothing$. The output of the r-approval fallback bargaining is the compromise set $CS_{d^*}^r(M)$ where d_r^* is $d_r^* = min\{d : CS_d^r(M) \neq \varnothing\}$.

We use fallback bargaining from [2] as a way for the agents to negotiate a judgment profile to replace their honest judgment profile and thus improve the judgment aggregation outcome.

Each agent $i \in N$ has a (truthfully) most preferred judgment set φ_i. Using Defintion 3.4, we construct a preference profile of total pre-orders over the elements of Ω. The matrix $M(m_{ij})$ will be of size $n \times t$ where t is the length of the longest preference order over profiles. The elements of the matrix m_{ij} are sets of judgment profiles which are in the same preference equivalence class for the agent i. The compromise set obtained with fallback bargaining over such constructed matrix M contains the judgment profiles which the agents agree on declaring for judgment aggregation.

Example 4.1 (DBAs matrix)**.** Consider the profile $\omega = (\varphi_1, \varphi_2, \varphi_3)$ containing the most preferred judgment sets for three agents introduced in Example 3.1. Having ω and using the H-score, we can construct the matrix M, given in Equation 8, over the universe Ω of judgment set profiles, i.e. "alternatives". Due to space limitations we will not list all the profiles contained in Ω and we use 111 to denote all judgment profiles for which f^{maj} yields the judgment set {1,1,1,} and 100,001,000 and $\lightning$ correspondingly for the remaining of the judgment profiles. For this example we will use the scoring which places $\lightning$ as the least preferred choice.

$$M^h = \left(\begin{array}{cc|cc} 111 & 100,010 & 000 & \lightning \\ 100 & 000 & 111,010 & \lightning \\ 010 & 000 & 111,100 & \lightning \end{array}\right) \quad (8)$$

For $r = 2$ a compromise set is reached for the depth $d_2^* = 2$ and $CS^2(M^h) = \{100, 000, 010\}$.
For $r = 3$, the depth of agreement is $d_2^* = 3$ with $CS^2(M^h) = \{100, 000, 010, 111\}$.

We present some properties of $CS_d^r(M)$. The proofs of the theorems are omitted due to space limitations.

Theorem 4.1. *Assume that $|\mathcal{A}| = k$. For a matrix M^h generated with the H-score, it holds that d_r^* of the reached agreement is with the following upper bound: $d_r^* \leq \lfloor \frac{(r-1)(k+1)+n}{n} \rfloor$*

This means that, as long as the agents always rank $\lightning$ as their least preferred alternative, $\lightning$ will not be an element of any $CS_d^r(M)$. In Example 4.1, d_r^* reaches its maximal value 3 for $r = n$, and again its maximal value $d_r^* = 2$ for $r = \lfloor \frac{n}{2} \rfloor + 1 = 2$.

It follows directly from the definition of fallback bargaining that any element in the compromise set will have a better utilitarian social welfare then an element not selected in the compromise set.

Theorem 4.2. *For any matrix M, aggregation function f and any scoring function $s : \Phi \longrightarrow (\Phi^{\lightning} \longrightarrow \mathbb{R})$ it holds that if $x \in CS_d^r(M)$, $d = n$ and $y \notin CS_d^r(M)$ then $USW(f(x)) < USW(f(y))$.*

This means that, if $\lightning$ is not a member of the compromise set, for every agent i there will exist an incentive for white manipulation. Consequently, the agents can chose any profile from the compromise set to be the declared profile for judgment aggregation. Clearly it is better to chose that profile $x \in CS_d^r(M)$ for which the $USW(f(x))$ is maximal. If $\lightning$ is in the compromise set, incentive for white manipulation can occur if there is another alternative $x \in CS_d^r(M)$ such that $USW(\lightning) \geq USW(f(x))$.

5 Conclusions and Future Work

The impossibility to avoid collectively irrational outcomes typically hampers judgment aggregation. Work in judgment aggregation has been focusing on developing procedures that avoid the impossibility by loosening aggregation conditions [7]. Instead, the present paper is looking into ways to incorporate the collectively irrational outcome. We incorporate the collectively irrational outcome by extending the standard judgment aggregation framework in a way that allows us to import techniques proper of social choice theory. Such extension has been obtained, along the lines followed in [3], by eliciting the preferences of the agents involved in the aggregation process via scoring functions. This has allowed us to introduce the notion of (utilitarian and egalitarian) social welfare in the judgment aggregation framework.

With this machinery in place, we initialize a study on a variant of the notion of manipulation of social choice rules which consider the welfare of the group—which we called "white manipulation". To the best of our knowledge, this is an unique attempt to present manipulation from a positive viewpoint. We have employed fallback bargaining [1] as an illustration of a procedure for realizing group white manipulation.

In our future work we intend to pursue the research lines presented here by taking into considerations the game-theoretic aspects of judgment aggregation in addition to the social-theoretic ones in scenarios which include the collectively irrational outcome. We also intend to refine the concept of white manipulation to capture, and study separately, the distinction between a group white manipulation and white manipulation conducted by a single agent.

Acknowledgments Davide Grossi is supported by *Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (VENI grant Nr. 639.021.816).

References

[1] S. Brams and D. Kilgour. Fallback bargaining. Working Papers 98-10, C.V. Starr Center for Applied Economics, New York University, 1998.

[2] S. Brams, D. Kilgour, and M. Sanver. A minimax procedure for negotiating multilateral treaties. In R. Avenhaus and I. Zartman, editors, *Diplomacy Games*, pages 265–282. Springer Berlin Heidelberg, May 2007.

[3] F. Dietrich and C. List. Strategy-proof judgment aggregation. STICERD - Political Economy and Public Policy Paper Series 09, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE, Aug 2005.

[4] W. Gaertner. *A Primer in Social Choice Theory*. Oxford University Press, 2006.

[5] L. Kornhauser and L. Sager. Unpacking the court. *Yale Law Journal*, 96:82–117, 1986.

[6] C. List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, 2002.

[7] C. List and C. Puppe. Judgment aggregation: A survey. In P. Anand, C. Puppe, and P. Pattanaik, editors, *Oxford Handbook of Rational and Social Choice*. Oxford University Press, forthcoming.

[8] P. Pettit. Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11:268–299, 2001.

[9] G. Pigozzi. Belief merging and the discursive dilemma: an argument-based account to paradoxes of judgment aggregation. *Synthese*, 152(2):285–298, 2006.

[10] A. D. Taylor. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press, 2005.